A THEORETICAL AND EMPIRICAL INVESTIGATION OF THE BOX-COX MODEL AND A NONLINEAR LEAST SQUARES ALTERNATIVE

Ernst R. Berndt
Mark H. Showalter
Jeffrey M. Wooldridge

Massachusetts Institute of Technology

WP#3187-90-EFA August 1990

A Theoretical and Empirical Investigation of the Box-Cox Model and a Nonlinear Least Squares Alternative

by

Ernst R. Berndt
Mark H. Showalter
Jeffrey M. Wooldridge*

*Professor of Applied Economics in the Sloan School at MIT, Ph.D. graduate student in the Sloan School, and Assistant Professor in the Department of Economics at MIT, respectively. Mssrs. Berndt and Showalter gratefully acknowledge the financial support of the Center for Energy Policy Research at MIT. We have benefited from the comments of participants at the MIT-Harvard joint econometrics seminar, at Boston College, and at the University of Montréal.

1. Introduction

The choice of an appropriate functional form is a very important issue in applied econometrics, for in many cases the underlying economic theory provides only limited guidance, e.g., use a functional form in which theoretically-consistent homogeneity and/or symmetry restrictions can be imposed. In practice, the Box-Cox and Box-Tidwell procedures have often been employed to choose among alternative functional forms. Spitzer (1984), however, has noted a very serious lack-of-scale invariance for t-statistics that emerges when one uses the Box-Cox or Box-Tidwell method. Moreover, unless the Box-Cox transformation parameter λ on the dependent variable y equals 0 or 1, one cannot solve for or compute $\hat{y}$, the fitted value of y, in closed form.

Recently, Wooldridge (1990) has developed an alternative to the Box-Cox procedure, one based on direct nonlinear least squares methods. With the nonlinear least squares (NLS) procedure, one can easily solve the fitted value problem inherent to the Box-Cox (BC) method, and although t-statistics on slope coefficients for the NLS estimator also lack invariance to arbitrary scaling of the dependent variable, Wooldridge has outlined how scale-invariant test statistics for exclusionary hypotheses can be conducted using the Lagrange multiplier test procedure. Moreover, since one important characteristic of the Box-Cox procedure is that it transforms the distribution of the dependent variable, Wooldridge has also derived computational formulae for obtaining heteroskedasticity-robust standard error estimates.

To understand better how important these theoretical issues might be in practice, it is necessary to implement the various procedures empirically and then to compare them. In this paper, therefore, we undertake an empirical comparison of the Box-Cox, nonlinear least squares, and weighted nonlinear least squares estimation procedures. Although choice of functional form is an important issue in almost all areas of applied econometrics, this issue is of special interest in labor economics and in hedonic pricing studies. We have therefore chosen one data set comparable to those commonly used in labor economics and two data sets previously employed in hedonic applications for our empirical comparisons.

The first data set, called CPS78, is similar to that used by many labor economists in estimating wage rate (or statistical earnings) equations and returns to education; this data set consists of 550 observations, randomly drawn from the May 1978 U.S. Current Population Survey. The second data set, called COLE, is that used by Cole et al. (1986) in their hedonic pricing study of mainframe computers in the U.S. from 1972 to 1984. This study is of special interest since in part it formed the basis of the official quality-adjusted price indexes for mainframe computers recently published by the U.S. Bureau of Economic Analysis.[1] The third data set, called CHOW, is that underlying the classic study by Chow (1967) of prices and the price elasticity of demand for mainframe computers in the U.S. from 1960 to 1965.

The outline of this paper is as follows. We begin with a theoretical overview, drawn in part from Wooldridge (1990). After providing a brief summary discussion of data sets and sources in section 3, in section 4 we present empirical evidence on the extent of scale invariance (or lack thereof), a comparison of parameter estimates and inference for the BC, NLS and WNLS estimators, and a comparison of residuals and goodness-of-fit measures. In section 5 we summarize, conclude, and outline future research issues.

2. Approaches to Generalizing Functional Form

The primary purpose of this paper is to compare competing methodologies for generalizing functional form in econometrics. Although there are several approaches to generalizing functional form, most can be put into one of two broad classes: (i) transformation methods, and (ii) methods that directly specify flexible functional forms for $E(y \mid x)$. The former

[1]See Cole et al. (1986) and Cartwright (1986).

category is dominated by the approach popularized by Box and Cox (1964), and its numerous extensions; for a survey, see Spitzer (1982). An approach which works directly with $E(y \mid x)$, which will simply be referred to as "nonlinear regression methods", has recently been suggested by Wooldridge (1990). In this section we discuss each of these approaches in turn, focusing on the issues of robustness, efficiency, and scale invariance. We also consider goodness-of-fit measures for choosing among the approaches in actual empirical applications.

2.1 Transformation Methods

Let $y > 0$ be the variable of interest, and let $x = (1, x_2, \ldots, x_K)$ be the vector of explanatory variables.[2] Throughout this paper, x can represent nonlinear transformations of an underlying set of variables; this is in contrast to y, which should be the economic variable of interest. Note that we assume **x** contains a constant.

Rather than postulating a model for $E(y \mid x)$ directly, the simplest transformation methods seek to find a transformation of y which has a linear conditional expectation. Although other classes could be considered, this paper focuses on the well-known Box-Cox transformation. For strictly positive y, define

$$(2.1) \qquad y(\lambda) \equiv \frac{y^{\lambda}-1}{\lambda}, \qquad \lambda \neq 0$$

$$(2.2) \qquad \equiv \log(y), \qquad \lambda = 0.$$

Given this definition, the weakest assumption employed in transformation methods is that, for some $\lambda \in \Re$ and some $K \times 1$ vector $\beta \in \Re^K$,

$$(2.3) \qquad E[y(\lambda) \mid x] = x\beta$$

[2]The Box-Cox procedure is ill defined for the case $y \leq 0$.

(throughout, the "true" values of lambda and beta are denoted λ and β). By itself, of course, assumption (2.3) is not enough to recover $E(y \mid x)$ as a function of x, β , and λ. This is not to say that λ and β cannot be estimated under (2.3); see, for example, Amemiya and Powell (1981).[3]

Because we take $E(y \mid x)$ to be of substantial interest, (2.3) in isolation is almost useless. While β_j measures the effect of x_j on $E(y(\lambda)|x)$, it generally tells us nothing about the effect of x_j on $E(y \mid x)$. As discussed in Wooldridge (1990), a consistent framework for defining elasticities, semi-elasticities, and other economic quantities generally requires these quantities to be defined in terms of $E(y \mid x)$. To move from $E(y(\lambda)|x)$ to $E(y \mid x)$, at least one additional distributional assumption on $D(y|x)$ is needed. One natural assumption is that log(y) is normally distributed with constant variance. Although plausible, this is not the preferred assumption in the literature, as the transformation $y(\lambda)$ is presumed to "regularize" the distribution in addition to yielding a linear conditional expectation. Thus, the original Box-Cox model assumes that, for some $\lambda \in \Re$, there exists $\beta \in \Re^K$ and $\sigma^2 > 0$ such that

$$(2.4) \qquad y(\lambda)\,|x \sim N(x\beta, \sigma^2)$$

(see also Spitzer (1982) and Hinkley and Runger (1984)). As has been observed by many statisticians and econometricians, (2.4) cannot strictly be true unless $\lambda = 0$. Thus, the distribution $D(y|x)$ -- and in particular the expectation $E(y \mid x)$ -- implied by (2.4) is not well-defined, nor do the traditional consistency properties of maximum likelihood estimation carry over (see, for example, Draper and Cox (1969) and Amemiya and Powell (1981)).

From our perspective, what is important is how well the approximation (2.4) allows one to estimate $E(y \mid x)$. This raises the rather important issue of how one obtains predictions of y in transformation models. Under (2.4), one might use the naive approach as suggested by

[3]However, Khazzoom (1989) has recently revealed some shortcomings of Amemiya and Powell's nonlinear 2SLS estimator in the Box-Cox context.

equations (2.5) and (2.6):

$$E(y|x) = [1+\lambda x\beta]^{1/\lambda},\ \lambda \neq 0 \tag{2.5}$$

$$= \exp(x\beta),\ \lambda = 0. \tag{2.6}$$

Of course these are incorrect given (2.4), but neither is there a well-defined conditional expectation function. However, one might want to use the normality assumption more intensively. If one believes that $u|x$ is distributed (approximately) as $N(0,\sigma^2)$, then the natural expectations are

$$E(y|x) = \int_{u \geq -\frac{1}{\lambda}(1+\lambda x\beta)}^{+\infty} [1+\lambda x\beta+\lambda u]^{1/\lambda} \frac{(1/\sigma)\phi(u/\sigma)}{\Phi(\frac{1}{\lambda\sigma}(1+\lambda x\beta))} du, \qquad \lambda > 0 \tag{2.7}$$

$$= \int_{-\infty}^{u \leq -\frac{1}{\lambda}(1+\lambda x\beta)} [1+\lambda x\beta+\lambda u]^{1/\lambda} \frac{(1/\sigma)\phi(u/\sigma)}{\Phi(-\frac{1}{\lambda\sigma}(1+\lambda x\beta))} du, \qquad \lambda < 0 \tag{2.8}$$

$$= \exp(\sigma^2/2+x\beta),\ \lambda = 0, \tag{2.9}$$

where $\phi(z)$ denotes the standard normal density and $\Phi(z)$ denotes the standard normal distribution function. The integration in (2.7) and (2.8) must be restricted to the regions $u \geq -\lambda^{-1}(1 + \lambda x\beta)$ and $u \leq -\lambda^{-1}(1 + \lambda x\beta)$, respectively, to ensure that the expectation is well-defined. Estimates of $E(y|x)$ are obtained from (2.7)-(2.9) once $\hat{\lambda}$, $\hat{\beta}$, and $\hat{\sigma}^2$ have been computed, usually by quasi-maximum likelihood methods.

Given observations $\{(x_t,y_t):t=1,2,\ldots,N\}$, the QMLE's of λ, β, and σ^2 solve the problem

$$(2.10) \qquad \max_{\lambda,\beta,\sigma^2} \sum_{t=1}^{N} \ell_t(\lambda,\beta,\sigma^2)$$

where $\ell_t(\lambda,\beta,\sigma^2)$ is the conditional log-likelihood of y_t given x_t for observation t:

$$(2.11) \qquad \ell_t(\lambda,\beta,\sigma^2) \equiv k_0 - \left(\frac{1}{2}\right)\log(\sigma^2) - \left(\frac{1}{2\sigma^2}\right)(y_t(\lambda) - x_t\beta)^2 + (\lambda-1)\log(y_t)$$

(in a time series context, it is assumed that $D(y_t|x_t) = D(y_t|x_t, \quad y_{t-1}, x_{t-1}, \ldots)$, i.e. there is no dynamic misspecification). Let $\theta \equiv (\lambda,\beta,\sigma^2)$ be the vector of parameters, and let

$$(2.12) \qquad s_t(\theta) \equiv \nabla_\theta \ell_t(\theta)$$

be the 1 x (K+2) gradient of the conditional log-likelihood with respect to θ. If the distributional assumption (2.4) (approximately) holds then the QMLE $\hat{\theta}$ is asymptotically normally distributed about θ with asymptotic variance estimated most easily by

$$(2.13) \qquad \hat{\mathfrak{S}}^{-1} \equiv \left(\sum_{t=1}^{N} s_t(\hat{\theta})' s_t(\hat{\theta})\right)^{-1}.$$

The asymptotic standard errors are obtained as the square roots of the diagonal elements of $\hat{\mathfrak{S}}^{-1}$. [4] If (2.4) fails to hold then $\hat{\theta}$ is generally inconsistent for θ. Nevertheless, Draper and Cox (1969) argue that $\hat{\theta}$ is approximately consistent provided the distribution of $y(\lambda)$ is symmetric.[5] When u|x depends on x, e.g. $\mathrm{Var}(y(\lambda)|x)$ is nonconstant, the QMLE based on the normality assumption (2.3) can be poorly behaved (e.g. Amemiya and Powell (1981); Seakes and Layson (1983)). In addition, Poirier (1978) finds that even if u is independent of x, the QMLE

---

[4]On this, however, see Amemiya and Powell (1981).

[5]Note that Draper and Cox also assume that u is independent of x, so that theirs is a very limited finding.

can exhibit severe asymptotic bias if $y(\lambda)$ is asymmetric. Generally speaking, the Box-Cox MLE can be very sensitive to the assumptions of homoskedasticity and normality of $y(\lambda)$. This is not a very desirable property of a method if it is intended primarily to generalize functional form.

In addition to it being nonrobust, the Box-Cox approach has another undesirable feature. This has to do with the lack of scale invariance of the t-statistics on $\hat{\beta}_j$, j=1,...,K. As was pointed out by Spitzer (1984) and others, the t-statistics on the coefficients $\hat{\beta}_j$, j=1,...,K, can be altered simply by multiplying y_t by a nonzero constant. This is unfortunate because the units of measurement of y is frequently arbitrary in economics (e.g. whether price is recorded in hundreds or thousands of dollars should be irrelevant). Dagenais and Dufour (1986) have noted that certain other nonlinear models have this feature (below, we show that the nonlinear regression approach also suffers from this problem).

One solution to this problem is, rather than to report t-statistics for coefficients, to use scale invariant likelihood ratio or Lagrange multiplier statistics for testing exclusion of each variable. For example, let $\tilde{s}_t \equiv s_t(\tilde{\theta}_{(j)})$ denote the 1 x (K+2) score evaluated at $\tilde{\theta}_{(j)}$, where $\tilde{\theta}_{(j)}$ is the QMLE computed under the restriction $\beta_j = 0$. Dagenais and Dufour (1986) show that the outer product LM statistic, obtained as $N - SSR = NR_u^2$ from the regression

$$1 \text{ on } \tilde{s}_t, \quad t=1,\ldots,N, \tag{2.14}$$

is invariant to the scale of y. Under H_0: $\beta_j = 0$, NR_u^2 is distributed approximately as χ_1^2 , provided (2.4) holds. While this is an attractive alternative, it is computationally expensive because it requires estimation of K distinct Box-Cox regression models (one for each β_j). Also, outer product forms of LM statistics are notorious for their poor finite sample properties; see Bollershev and Wooldridge (1988).

Rather than use an LM statistic, Dagenais and Dufour suggest using a particular version of Neyman's (1959) $C(\alpha)$ statistic. For the Box-Cox model, this results in significant computational advantages because λ need only be estimated once, from the unrestricted model.

The reader is referred to the Dagenais and Dufour (1986) paper for further details.

Pedagogically, the LM and $C(\alpha)$ approaches are unsatisfying because they do not allow construction of confidence intervals; hypotheses of the form H_0: $\beta_j = b_j$ using an LM or $C(\alpha)$ test require a separate computation whenever b_j changes. Researchers typically prefer to see standard errors attached to parameter estimates, so that a confidence interval can be constructed, and t-statistics (Wald statistics) can be computed for testing any hypothesis of the form H_0: $\beta_j = b_j$. Thus, it seems useful to attempt to salvage the usual t-statistic.

The problem with the usual t-statistic is that, as the analyst searches over different scalings of y, the t-statistics of the $\hat{\beta}_j$ can be changed, sometimes (as we shall show) dramatically. This is, of course, a form of data mining, and is not attractive given the usual assumptions underlying statistical inference. One solution to this problem is to force the analyst to estimate a scale parameter using the sample data. This is what Spitzer (1984) recommends in order to obtain scale-invariant t-statistics. He suggests dividing each observation y_t by the sample geometric mean of $\{y_t$:t=1,...,N$\}$, say $\hat{v}$. The new variable $y_t/\hat{v}$ is trivially scale invariant, so one might think that this solves the problem.[6]

However, there are two potential problems with this approach. The first is that the coefficients $\hat{\beta}_j$ might become more difficult to interpret. Fortunately, this turns out not to be much of a problem because the estimate of marginal effects from (2.5) or (2.7) - (2.9) are affected exactly as one would expect: they are simply scaled down by $\hat{v}$. Moreover, point estimates of elasticities and semi-elasticities are invariant to the scaling of y.

The second problem is potentially more serious: one needs to account for the randomness of $\hat{v}$ when computing standard errors of $\hat{\beta}_j$ and $\hat{\lambda}$. The simplest way to address this problem is to view the estimator $\hat{\theta}$ as a two-step estimator. Thus consider the extended model

[6]One additional advantage of employing this geometric mean transformation is that with the transformed data, maximizing the log-likelihood function is equivalent to minimizing the sum of squared residuals. For a discussion of this computational nuance, see Zarembka (1968).

(2.15) $$y(\nu,\lambda) \sim N(x\beta,\sigma^2)$$

~~(2.16)~~ $$\nu \equiv \exp[E(\log(y))]$$

where $y(\nu,\lambda) = ((y/\nu)^{\lambda} - 1)/\lambda$, $\lambda \neq 0$, $y(\nu,\lambda) = \log(y/\nu)$, $\lambda = 0$. The parameter ν is the population geometric mean of y. Admittedly, ν could be defined to be one of a variety of other scale parameters, e.g. $\nu = E(y)$. A researcher is free to choose any moment of y, provided the moment exists. The important point is that, no matter how ν is defined, because it is estimated using sample data, the variance of the estimator $\hat{\nu}$ should be accounted for in any inference procedures.

The QMLE $\hat{\theta}$ now solves[7]

(2.17) $$\max_{\theta} \sum_{t=1}^{N} \ell_t(\theta;\hat{\nu}), \quad \text{where} \quad \hat{\nu} \equiv \exp\left(\frac{1}{N}\sum_{t=1}^{N} \log(y_t)\right) \quad \text{and}$$

(2.18) $$\ell_t(\theta,\nu) \equiv k_0 - (1/2)\log(\sigma^2) - (1/2)(y_t(\nu,\lambda) - x_t\beta)^2/\sigma^2 + (\lambda-1)\log(y_t/\nu).$$

A standard mean value expansion can be used to derive an estimate of the asymptotic variance of $\hat{\theta}$. Redefining s_t to account for the dependence upon η gives $s_t(\theta,\nu) \equiv \nabla_{\theta}\ell_t(\theta,\nu)$, the 1x(K+2) score of the log-likelihood for observation t. Let $\hat{s}_t \equiv s_t(\hat{\theta},\hat{\nu})$ equal the score evaluated at the estimated values $\hat{\theta},\hat{\nu}$. Now define $\hat{g}_t \equiv \hat{s}_t + \hat{C}_N \cdot \hat{\nu} \cdot \log(y_t/\hat{\nu})$ where $\hat{C}_N \equiv \frac{1}{N} \cdot \sum_{t=1}^{N} \nabla_{\nu}\hat{s}'_t$. Then, as can be shown using methods similar to those employed in the

[7]Alternatively, one might simply want arbitrarily to choose one observation as "numeraire". While such a procedure might be "natural" in a time series context (say, take the first observation), in a cross-section context the choice of numeraire observation would seem to be totally arbitrary, yielding somewhat capricious t-statistics.

appendix to Wooldridge (1990), a consistent estimator of the asymptotic variance of $\hat{\theta}$ is

(2.19)
$$\left(\sum_{t=1}^{N} \hat{s}_t'\hat{s}_t\right)^{-1}\left(\sum_{t=1}^{N} \hat{g}_t'\hat{g}_t\right)\left(\sum_{t=1}^{N} \hat{s}_t'\hat{s}_t\right)^{-1}.$$

Note that if the estimation of $\hat{v}$ were ignored, this reduces to the usual outer product of the score estimator. Surprisingly, the standard error of λ, $se(\hat{\lambda})$, obtained from (2.19), differs from that obtained from (2.13), even though $se(\hat{\lambda})$ is scale invariant for any fixed scaling of y.

2.2 The Nonlinear Least Squares Approach

Let y and $x = (1,x_2,...,x_k)$ be defined as in the previous subsection. Without any assumptions on the conditional distribution of y given x (except that its support is contained in $[0,\infty)$), consider the following model for $E(y\,|\,x)$:

(2.20)
$$E(y|x) = [1+\lambda x\beta]^{1/\lambda},\ \lambda\neq 0$$

(2.21)
$$= \exp(x\beta),\ \lambda=0.$$

When $\lambda = 1$ (2.20) reduces to a linear model for $E(y\,|\,x)$.[8] The exponential regression model (2.21) is particularly appealing for a strictly positive y because it ensures that the predicted values are well-defined and positive for all x and any value of β, whereas this is not necessarily the case for (2.20). Note that the semi-elasticity for this model is

[8]Interestingly, the conditional mean functions (2.20) and (2.21) can be derived from a modified version of the Box-Cox model if $P(y > 0) = 1$. If the conditional mean assumption (2.3) is supplemented with the assumption that $\log(y)|x$ is normally distributed with constant variance, then (2.20) and (2.21) can be shown to hold; see Wooldridge (1990) for details.

(2.22) $$\frac{\partial E(y|x)}{\partial x_j}\cdot\frac{1}{E(y|x)}=[1+\lambda x\beta]^{-1}\beta_j,$$

while the elasticity is

(2.23) $$\frac{\partial E(y|x)}{\partial x_j}\cdot\frac{x_j}{E(y|x)}=[1+\lambda x\beta]^{-1}\beta_j x_j,$$

To estimate β and λ by nonlinear least squares (NLS) or weighted NLS (WNLS), the derivatives of the regression function are needed. Define the (K+1)x1 parameter vector $\delta = (\beta,\lambda)'$ and express the parameterized regression function for $E(y|x)$ as

(2.24) $$\begin{aligned} m(x;\delta) &= [1+\lambda x\beta]^{1/\lambda}, && \lambda\neq 0 \\ &= \exp(x\beta), && \lambda=0. \end{aligned}$$

For $\lambda \neq 0$ the gradient of $m(x;\delta)$ with respect to β is the 1xK vector

(2.25) $$\nabla_\beta m(x;\delta)=[1+\lambda x\beta]^{(1/\lambda)-1}x.$$

For $\lambda = 0$,

(2.26) $$\nabla_\beta m(x;\beta,0)=\exp(x\beta)x.$$

The derivative of $m(x;\delta)$ with respect to λ, when $\lambda \neq 0$, is derived in Wooldridge (1990) as

(2.27) $$\nabla_\lambda m(x;\beta,\lambda)=\frac{1}{\lambda^2}[1+\lambda x\beta]^{\frac{(1-\lambda)}{\lambda}}[\lambda x\beta-(1+\lambda x\beta)\log(1+\lambda x\beta)].$$

For $\lambda = 0$ it equals

(2.28) $$\nabla_\lambda m(x;\beta,0)= -\frac{\exp(x\beta)(x\beta)^2}{2}$$

Equation (2.27) is the basis for the LM statistic for the hypothesis H_0: $\lambda = \lambda_0$, while (2.28) is the basis for the LM test of H_0: $\lambda = 0$; see Wooldridge (1990) for further details.

Under the assumption that (2.24) holds, δ can be consistently estimated by NLS. In addition, if $V(y|x)$ is constant and equal to, say η^2, then standard formulae are available for estimating the asymptotic variance of the NLS estimator. Let $\hat{\delta}$ be the NLS estimator, let $\hat{e}_t \equiv y_t - m(x_t, \hat{\delta})$ denote the NLS residuals, and estimate η^2 by the degrees-of-freedom adjusted estimator

$$\hat{\eta}^2 \equiv \frac{1}{N-P} \sum_{t=1}^{N} \hat{e}_t^2, \tag{2.29}$$

where $P = K+1$ is the number of parameters. A standard estimate of the asymptotic variance of $\hat{\delta}$ is

$$\hat{\eta}^2 \left(\sum_{t=1}^{N} \nabla_\delta \hat{m}'_t \nabla_\delta \hat{m}_t \right)^{-1} \tag{2.30}$$

which is valid provided that $V(y_t|x_t) = \eta^2$ and, in a time series context, $E(y_t|x_t) = E(y_t|x_t, y_{t-1}, x_{t-1}, \ldots)$. (This latter condition ensures that the errors $e_t = y_t - E(y_t|x_t)$ are conditionally serially uncorrelated.)

The heteroskedasticity-robust asymptotic variance estimator of β and λ can be obtained by using the approach of White (1980). That estimator in this context is

$$\frac{N}{(N-P)} \left(\sum_{t=1}^{N} \nabla_\delta \hat{m}'_t \nabla_\delta \hat{m}_t \right)^{-1} \left(\sum_{t=1}^{N} \hat{e}_t^2 \nabla_\delta \hat{m}'_t \nabla_\delta \hat{m}_t \right) \left(\sum_{t=1}^{N} \nabla_\delta \hat{m}'_t \nabla_\delta \hat{m}_t \right)^{-1}, \tag{2.31}$$

which uses a degrees-of-freedom adjustment to enhance finite sample performance.

Although the NLS estimator is robust to heteroskedasticity and requires no distributional assumption, the estimate in (2.31) might be large if $V(y_t|x_t)$ is highly variable. Improvements in efficiency might be realized by using a weighted NLS approach. Let $\hat{\omega}_t \equiv \omega(x_t, \hat{\gamma})$ be a set of weights that can depend on x_t and a vector of estimates $\hat{\gamma}$. Then, if

$$V(y_t|x_t) = \eta^2 \omega(x_t, \gamma) \tag{2.32}$$

and $\hat{\gamma}$ is $\sqrt{N}$-consistent for γ, η^2 is estimated as in (2.29) where $\hat{e}_t$ is now weighted by $1/\sqrt{\hat{\omega}_t}$. An estimator of the asymptotic variance of the WNLS estimator $\hat{\delta}$ is given by (2.30), except that $\hat{e}_t$ and $\nabla_\delta \hat{m}_t$ are weighted by $1/\sqrt{\hat{\omega}_t}$. In the empirical work, we use the NLS estimator and a WNLS estimator with

$$\omega(x, v) = [m(x, \delta)]^2; \tag{2.33}$$

the estimated $\hat{\omega}_t$ weights are obtained as the squared fitted values from the initial NLS regression.

The Lagrange multiplier tests for the linear (λ=1) and exponential(λ=0) models are simple to compute. In the general WNLS case, the test for H_0: $\lambda = 1$ is obtained as NR_u^2 from the regression

$$\frac{\hat{e}_t}{\sqrt{\hat{\omega}_t}} \ on \ \frac{x_t}{\sqrt{\hat{\omega}_t}}, \frac{\hat{y}_t \log(\hat{y}_t)}{\sqrt{\hat{\omega}_t}} \quad t=1,\dots,N, \tag{2.34}$$

where $\hat{y}_t$ and $\hat{e}_t$ are the fitted values and residuals from the WNLS estimation. LM is distributed as χ_1^2 under H_0 if the variance assumption (2.33) holds. To obtain a robust form of the statistic, first compute the residuals $\tilde{r}_t$ from the regression

$$\frac{\hat{y}_t}{\sqrt{\hat{\omega}_t}} \log(\hat{y}_t) \ on \ \frac{x_t}{\sqrt{\hat{\omega}_t}}, \tag{2.35}$$

and then form LM = N - SSR from the regression

$$1 \ on \ \tilde{e}_t \cdot \tilde{r}_t, \ t=1,\dots,N, \tag{2.36}$$

where $\tilde{e}_t \equiv \hat{e}_t/\sqrt{\hat{\omega}_t}$. Testing H_0: $\lambda = 0$ requires WNLS estimation of an exponential regression model. Let $\hat{y}_t$ and $\hat{e}_t$ be the fitted values, and let $\tilde{y}_t$ and $\tilde{e}_t$ be the weighted quantities. Then, from Wooldridge (1990), compute $LM = NR_u^2$ from the OLS regression

$$(2.37) \qquad \tilde{e}_t \text{ on } \tilde{y}_t \cdot x_t,\ \tilde{y}_t \cdot [\log(\hat{y}_t)]^2;$$

again, LM $\overset{a}{\sim} \chi_1^2$ under H_0 and (2.33). The robust test is based on the residuals, $\tilde{r}_t$, from the OLS regression

$$(2.38) \qquad \tilde{y}_t [\log(\hat{y}_t)]^2 \text{ on } \tilde{y}_t \cdot x_t,$$

and the LM test statistic is then computed as LM = N - SSR, as in (2.36).

One problem that the NLS and WNLS approaches share with transformation methods is that the t-statistics of $\hat{\beta}_1, \ldots, \hat{\beta}_K$ lack invariance to the scaling of y whenever λ is estimated along with the β's. Wooldridge (1990) shows that the estimate and standard error for λ (both the usual and robust form) are invariant to rescaling of y, and moreover, that the LM statistic for testing any exclusion restrictions on the β_k slope coefficients is also invariant to the scaling of y; hence LM tests can be used as alternatives to t-statistics. Unfortunately, if there are many x's, this can be computationally expensive.

As one method of obtaining scale invariant t-statistics, consider adding a scale parameter to (2.24):

$$(2.39) \qquad \begin{aligned} &m(x_t, \beta, \lambda, \nu) = \nu [1 + \lambda x_t \beta]^{\frac{1}{\lambda}}, \quad \lambda \neq 0, \\ &m(x_t, \beta, \lambda, \nu) = \nu \exp(x_t \beta), \quad \lambda = 0, \\ &\nu \equiv \exp[E(\log(y_t))]; \end{aligned}$$

again, this definition of η is arbitrary, but it is estimable from the data. Let $\hat{\nu}$ be the sample geometric mean of $\{y_t\}$. Then substituting $\hat{\nu}$ into (2.39) and using NLS or WNLS is the same as dividing each y_t by $\hat{\nu}$ and estimating the model as before. As in the Box-Cox case, the variation in $\hat{\nu}$ must be taken into account. Wooldridge (1990) has derived a consistent estimator of the asymptotic variance of $\hat{\delta}$ where $\hat{\delta}$ now refers to the set of parameters in the (sample geometric mean scaled) model.

Let

$$(2.40) \qquad \tilde{C}_N \equiv \frac{1}{N}\sum_{t=1}^{N}\left(\frac{\nabla_\delta \hat{m}_t}{\sqrt{\hat{\omega}_t}}\right)'\left(\frac{\nabla_v \hat{m}_t}{\sqrt{\hat{\omega}_t}}\right),$$

where $\nabla_\delta \hat{m}_t$ is the same as before except that it is not multiplied by $\hat{\vartheta}$; also, note that $\nabla_v \hat{m}_t = [1+\hat{\lambda}x\hat{\beta}]^{1/\hat{\lambda}}$ is simply the fitted value for the scaled regressand $y_t/\hat{\vartheta}$. A consistent estimate of the asymptotic variance of $\hat{\theta}$ is

$$(2.41) \qquad \left(\sum_{t=1}^{N} \nabla_\delta \tilde{m}'_t \nabla_\delta \tilde{m}_t\right)^{-1} [I_P|-\tilde{C}_N]\left(\sum_{t=1}^{N} \tilde{g}'_t \tilde{g}_t\right)[I_P|-\tilde{C}_N]'\left(\sum_{t=1}^{N} \nabla_\delta \tilde{m}'_t \nabla_\delta \tilde{m}_t\right)^{-1}$$

where $P \equiv K+1$, $\tilde{g}_t$ is the $1x(P+1)$ vector

$$(2.42) \qquad \tilde{g}_t \equiv \left(\tilde{e}_t \nabla_\delta \tilde{m}_t,\ \hat{\vartheta}\log\left(\frac{y_t}{\hat{\vartheta}}\right)\right)$$

and all "~" variables are weighted by $1/\sqrt{\hat{\omega}_t}$. This expression is robust to variance misspecification and also accounts for the randomness of $\hat{\vartheta}$.

2.3 Goodness-of-Fit Measures

From our perspective, the ultimate goal of any exercise to generalize functional form is to obtain reliable estimates of $E(y|x)$. Therefore, it is important to have goodness of fit measures that allow discrimination among alternative methods. One natural measure is simply an R-squared defined in terms of the untransformed variable y. Given fitted values $\hat{y}_t$, $t=1,\dots,N$, the R-squared is simply

$$(2.43) \qquad R^2 \equiv 1-\frac{\sum_{t=1}^{N}(y_t-\hat{y}_t)^2}{\sum_{t=1}^{N}(y_t-\bar{y})^2}.$$

This R-squared measures the percentage of the variation in y "explained by" the x's, regardless of how the fitted values $\hat{y}_t$ are obtained. In the context of NLS or WNLS, $\hat{y}_t$ is obtained directly as $\hat{y}_t \equiv m(x_t, \hat{\delta})$. The fitted values from a transformation model can be obtained only once an expression for E(y | x) is available. In the Box-Cox context, this expression is given by (2.7)-(2.9).

The R-squared defined by (2.43) is not without its problems. First, for a given functional form for E(y | x), R^2 is always maximized by the NLS estimator. Consequently, (2.43) cannot be used to choose between weighted and unweighted least squares estimators. This is less of a problem that it might seem because, provided E(y | x) is correctly specified, the fitted values from these procedures should be similar.

In this paper, the primary use of R^2 is to compare the direct procedures, NLS and WNLS, to transformation methods (specifically Box-Cox). Because the models for E(y | x) are nonnested for these two approaches, R^2 can legitimately be used as a goodness-of-fit measure. However, since transformation methods do not directly minimize the sum of squared residuals in y_t, R^2 criteria will tend to favor NLS.

An alternative to a standard R-squared measure would be a metric of how well the models estimated the conditional variance as well as the conditional mean. One possibility is simply to evaluate a normal log-likelihood function at the implied conditional mean and conditional variance. It is well-known that the true conditional mean and the true conditional variance maximize the expected log-likelihood whether or not y conditional on x is normally distributed. We do not pursue this approach for two reasons. First, estimation of V(y|x) in the Box-Cox context is computationally cumbersome (it is more difficult than estimating E(y|x)). Second, a primary motivation for using NLS and WNLS is that the estimates are robust to heteroskedasticity in the former case and variance misspecification in the latter case. A goodness-of-fit measure based on the conditional mean and conditional variance essentially ignores the robustness considerations.

In addition to aggregate R^2 statistics which summarize the sample information in a single statistic, one might be interested in how well individual order statistics of the various empirical distributions of the fitted values match with those of the dependent variable. While this does not provide a metric for comparison across samples, it should provide clues to possible anomolies in the estimation procedures.

In a similar vein we can calculate the correlation matrices of the various fitted values and the dependent variables. This gives a direct measure of how well the various procedures match the dependent variable and also what similarities exist among the estimation procedures.

In the empirical section of the paper we shall use several of these variations to assess goodness-of-fit for each of the sample data sets.

3. Data and Empirical Implementation

For the BC procedure, we compute standard error estimates employing the variance-covariance matrix of analytic first derivatives as outlined by Berndt, Hall, Hall and Hausman (1974), and the estimate corrected for the sample geometric mean (see (2.19)). For NLS and WNLS, we compute three sets of standard errors using: first, the Gaussian quadratic form of analytic first derivatives; second, a heteroskedasticity-robust estimator due to White (1980); and third, a heteroskedasticity-robust estimator which also accounts for the estimated geometric mean of y (equation (2.31)).

The data sets used for comparing the alternative estimators have been chosen to generate additional interest in classic findings and to facilitate replication, and are all taken from the data diskette accompanying Berndt (1990). Specifically, we employ three data sets. The first, called CPS78, is a random sample of 550 observations drawn from the May 1978 U.S. Current Population Survey, originally constructed by Henry S. Farber. This type of data set is frequently employed by labor economists to estimate statistical earnings functions, where the dependent variable is some transformation (often logarithmic) of the hourly wage rate in dollars (WAGE), and the set of regressors includes a constant term, potential experience (EXP) (measured as age minus years of education minus schooling minus six), and its square (EXP2), a race dummy variable (RACE) taking on the value one only if the individual is non-white and non-Hispanic, a gender dummy variable (FEMALE) taking on the value one only if the individual is female, and an education variable (EDN) measuring the years of schooling.

A second data set is that underlying the classic study of quality-adjusted mainframe computer prices and the demand for computers by Gregory Chow (1967). Chow related the monthly rental price of mainframe computers (PRICE) to multiplication speed (MULT), memory capacity (MEM), access time (ACCESS), and a set of annual dummy variables; we employ Chow's 1961-1965 data, whose details are discussed further in Berndt (1990,ch.4).

The third and final data set we employ is that underlying the more recent mainframe computer hard disk drive price index study by Rosanne Cole et al. (1986). This study is of interest since it has played a critical role in the U.S. Bureau of Economic Analysis decision to employ hedonic regression methods to adjust mainframe computer prices for quality change over time in its official computer price index. The Cole et al. data set encompasses 91 models over the 1972-84 time period, and relates the list price of hard disk drives (PRICE) to a constant, 1973-84 annual time dummy variables, and two performance variables, SPEED and CAP (capacity); further details on these data are given in Cole et al. (1986), Triplett (1989) and Berndt (1990,Ch. 4).

Computations for this empirical research were carried out on an IBM 4381 mainframe and an AT&T 6386 personal computer, using the statistical programs TSP and GAUSS.

4. Empirical Results

Our discussion of empirical results focuses on three issues: (a) we begin by addressing the lack of scale invariance issue, assessing its numerical significance on estimated slope coefficients, and then we implement scale invariant t-statistics using the LM test procedure for both the BC and NLS estimators; (b) we then go on to discuss similarities and differences among the BC, NLS and WNLS parameter estimates in our three data sets, as well as the estimated standard errors; and finally (c) we compare the alternative estimation procedures using a variety of goodness-of-fit criteria.

4.1 Resolving the Lack of Invariance to Scaling Issue

We begin by demonstrating empirically, in a rather persuasive manner, the lack of invariance of t-statistics on slope coefficients to arbitrary re-scaling of the dependent variable in the BC and NLS estimation procedures, and then we present LM-based scale invariant t-statistics. Using the CPS78 data we estimated a model in which WAGE was measured in dollars per hour, and then multiplied this by 100, resulting in a measure in units of cents per hour. Results from the various estimations are presented in Table 1.

With both the BC and NLS procedures, as expected, estimates of λ and its standard error are invariant to scaling; the BC parameter estimate (standard error) is 0.072 (0.037), while that for NLS is -0.275(0.300). Note that λ=0 cannot be rejected at usual significance levels, thereby lending support to the common procedure in labor economics of employing log (WAGE) as the dependent variable in statistical earnings functions.

Matters are rather different, however, when we examine estimated slope coefficients and their associated t-statistics. For the BC estimator, although the RACE coefficient and t-statistic are relatively robust under scaling, other coefficients and t-statistic vary considerably, with some t-statistics changing by a factor of more than two. For the NLS estimates, this lack of robustness also is present; the coefficients and t-statistics on the EDN and EXP variables, for example, change by a factor greater than three after arbitrary re-scaling. Notice that no sign changes occur under re-scaling for either the BC or NLS estimators. Finally, for each coefficient in the line labeled "LM t-Stat", we present NLS scale-invariant t-statistics based on the LM test statistic, computed as the (positive) square root of the χ_1^2 test statistic based on (2.35) with weights all equal to one. We conclude that while the issue of scaling is empirically significant for these models in particular, and for nonlinear models in general, scale-invariant t-statistics can be obtained using the LM test procedure.

--- Table 1 ---

Resolution of the Lack of Invariance to Scaling Issue

CPS 1978 Data

	Box-Cox		NLS	
	Unscaled ($ per hour)	Scaled (*100)	Unscaled ($ per hour)	Scaled (*100)
Lambda	0.072	0.072	-0.275	-0.275
Std Err	0.037	0.037	0.300	0.300
t-Stat	1.917	1.917	-0.917	-0.917
LM t-Stat	0.816	0.816	0.563	0.563
Constant	0.540	6.204	0.702	2.807
Std Err	0.120	0.596	0.099	1.685
t-Stat	4.515	10.412	7.087	1.666
LM t-Stat	9.734	9.734	4.844	4.844
FEMALE	-0.371	-0.515	-0.200	-0.056
Std Err	0.046	0.134	0.115	0.109
t-Stat	-8.054	-3.849	-1.741	-0.514
LM t-Stat	9.092	9.092	6.890	6.890
RACE	-0.131	-0.182	-0.094	-0.026
Std Err	0.050	0.075	0.063	0.052
t-Stat	-2.602	-2.420	-1.490	-0.507
LM t-Stat	2.485	2.485	2.236	2.236
EDN	0.080	0.111	0.042	0.012
Std Err	0.010	0.029	0.026	0.024
t-Stat	8.308	3.822	1.614	0.501
LM t-Stat	9.152	9.152	5.596	5.596
EXP	0.034	0.047	0.020	0.006
Std Err	0.007	0.015	0.012	0.011
t-Stat	4.734	3.078	1.690	0.511
LM t-Stat	6.412	6.412	5.695	5.695
EXP2	-4.10E-04	-5.69E-04	-2.35E-04	-6.62E-05
Std Err	1.40E-04	2.40E-04	1.47E-04	1.29E-04
t-Stat	-2.928	-2.377	-1.599	-0.513
LM t-Stat	3.777	3.777	3.173	3.173

Notes: The LM t-Stats are based on the LM test statistic (computed as the square root of the chi-squared test statistic), allowing for heteroskedasticity; they are not computed as the ratio of the parameter estimate to its standard error.

4.2 Comparison of Parameters and Standard Errors for BC, NLS and WNLS

Having disposed of the scaling issue for t-statistics, we now compare parameter estimates and standard errors for the BC, NLS and WNLS estimators, where in each case we follow Spitzer's (1984) suggestion and transform the dependent variable by the sample geometric mean, as discussed underneath (2.36). We begin with a comparison based on the CPS78 data discussed briefly in the previous sub-section.

As seen in Table 2, although estimates of λ vary in sign based on the BC, NLS and WNLS methods, standard errors are relatively large (especially when corrected for both heteroskedasticity and the random geometric sample mean), and it is not clear these estimates differ significantly. Coefficient estimates on the EDN, RACE, FEMALE ,EXP and EXP2 variables are also very similar across the BC, NLS and WNLS estimation procedures with this data set, but both usual, robust and corrected (for the geometric mean) NLS standard errors are larger than those for the BC and WNLS. The NLS t-statistics are not always largest, however, owing to variation in parameter estimates among estimations; scale-invariant LM t-statistics bear no systematic inequality relationship to the various values based on traditional computations.

The similarity in parameter estimates and inference obtained using the BC, NLS and WNLS procedures does not occur, however, for the COLE data set. As seen in Table 3, estimates of λ based on NLS (2.60) and WNLS (4.46) differ dramatically from that based on BC (0.87). The coefficients for SPEED also differ widely across models with (13.26) for BC, (10.27) for NLS and (-0.85) for WNLS. Only the WNLS estimate would be considered insignificantly different from zero. The estimates fo CAP are all positive and small. With respect to the annual time dummy variable coefficients estimates (used as a basis for forming quality-adjusted computer price indexes in the hedonic price literature), sign differences among BC, NLS and WNLS occur for four of the twelve coefficients--1973, 1974, 1977 and 1978. Interestingly,

--- Table 2 ---

Alternative Estimates of Standard Errors and t-Statistics
CPS 1978 Data

	BC			NLS			WNLS		
		Std.Err.	t-stat		Std.Err.	t-stat		Std.Err.	t-stat
Lambda	0.07151			-0.27532			-0.25248		
Usual		0.03730	1.917		0.30020	-0.917		0.30592	-0.825
Robust					0.45707	-0.602		0.30563	-0.826
Correct		0.03730	1.917		0.45707	-0.602		0.30563	-0.826
LM t-Stat			0.816			0.563			0.265
Constant	-1.10529			-1.02175			-1.08120		
Usual		0.10211	-10.825		0.14443	-7.074		0.10850	-9.965
Robust					0.11650	-8.771		0.10329	-10.467
Correct		0.10633	-10.395		0.11869	-8.608		0.10593	-10.206
LM t-Stat			9.734			4.844			6.599
EDN	0.07109			0.06712			0.07051		
Usual		0.00692	10.269		0.00950	7.062		0.00766	9.200
Robust					0.00805	8.336		0.00676	10.425
Correct		0.00692	10.268		0.00807	8.317		0.00679	10.389
LM t-Stat			9.152			6.890			6.950
RACE	-0.11591			-0.14922			-0.11212		
Usual		0.04548	-2.548		0.05757	-2.592		0.04684	-2.394
Robust					0.05113	-2.919		0.04856	-2.309
Correct		0.04549	-2.548		0.05111	-2.920		0.04857	-2.308
LM t-Stat			2.485			2.236			1.852
FEMALE	-0.32857			-0.31723			-0.30226		
Usual		0.03461	-9.492		0.04589	-6.913		0.03675	-8.225
Robust					0.04456	-7.119		0.03637	-8.311
Correct		0.03463	-9.487		0.04433	-7.156		0.03621	-8.347
LM t-Stat			9.092			5.596			7.587
EXP	0.02980			0.03115			0.03224		
Usual		0.00574	5.194		0.00583	5.348		0.00501	6.430
Robust					0.00574	5.426		0.00427	7.559
Correct		0.00574	5.192		0.00575	5.419		0.00429	7.523
LM t-Stat			6.412			5.695			6.410
EXP2	-0.00036			-0.00037			-0.00040		
Usual		0.00012	-3.024		0.00012	-3.204		0.00011	-3.717
Robust					0.00013	-2.957		0.00009	-4.215
Correct		0.00012	-3.024		0.00013	-2.957		0.00009	-4.209
LM t-Stat			3.777			3.173			3.868

--- Table 3 ---
Alternative Estimates of Standard Errors and t-Statistics
Cole Data

	BC			NLS			WNLS		
		Std.Err.	t-stat		Std.Err.	t-stat		Std.Err.	t-stat
Lambda	0.87037			2.59868			4.46257		
Usual		0.25618	3.397		0.38500	6.750		0.59394	7.513
Robust					0.39778	6.533		0.54324	8.215
Correct		0.25618	3.397		0.39778	6.533		0.54324	8.215
LM t-Stat			5.708			4.654			4.381
Constant	-0.63805			-0.55542			-0.30151		
Usual		0.10407	-6.131		0.07714	-7.200		0.02700	-11.165
Robust					0.08339	-6.660		0.02220	-13.583
Correct		0.11859	-5.380		0.08449	-6.574		0.02423	-12.446
LM t-Stat			5.412			2.846			2.561
SPEED	13.25730			10.27272			-0.85005		
Usual		2.55902	5.181		3.49064	2.943		0.83596	-1.017
Robust					4.12136	2.493		0.55766	-1.524
Correct		2.42761	5.461		4.08452	2.515		0.60237	-1.411
LM t-Stat			5.105			1.624			1.065
CAP	0.00015			0.00130			0.00329		
Usual		0.00016	0.914		0.00047	2.751		0.00059	5.533
Robust					0.00043	2.982		0.00066	4.964
Correct		0.00017	0.904		0.00048	2.697		0.00093	3.540
LM t-Stat			0.701			2.622			2.212
1973	0.23395			0.07420			-0.13808		
Usual		0.11844	1.975		0.11876	0.625		0.07631	-1.809
Robust					0.12153	0.611		0.07085	-1.949
Correct		0.11830	1.978		0.12150	0.611		0.07723	-1.788
LM t-Stat			1.428			0.657			1.183
1974	0.03488			-0.07558			-0.19763		
Usual		0.09998	0.349		0.08422	-0.897		0.05016	-3.940
Robust					0.08957	-0.844		0.06199	-3.188
Correct		0.09989	0.349		0.09075	-0.833		0.07636	-2.588
LM t-Stat			0.202			0.680			0.775
1975	-0.06817			-0.12804			-0.21171		
Usual		0.14017	-0.486		0.06795	-1.884		0.04544	-4.659
Robust					0.05547	-2.308		0.05243	-4.038
Correct		0.14027	-0.486		0.05609	-2.283		0.06796	-3.115
LM t-Stat			0.544			1.603			1.517
1976	-0.06437			-0.15052			-0.21393		
Usual		0.10070	-0.639		0.06054	-2.486		0.04477	-4.778
Robust					0.05144	-2.926		0.05209	-4.107
Correct		0.10092	-0.638		0.05226	-2.880		0.06800	-3.146
LM t-Stat			0.444			2.048			1.771
1977	0.02898			-0.11830			-0.17827		
Usual		0.11231	0.258		0.09211	-1.284		0.06020	-2.962
Robust					0.07356	-1.608		0.05690	-3.133
Correct		0.11223	0.258		0.07155	-1.653		0.06522	-2.733
LM t-Stat			0.192			1.701			2.128
1978	0.21194			-0.19218			-0.61490		
Usual		0.18925	1.120		0.28046	-0.685		0.69631	-0.883
Robust					0.22982	-0.836		0.42479	-1.448
Correct		0.18835	1.125		0.22799	-0.843		0.41363	-1.487
LM t-Stat			1.332			0.964			1.445
1979	-0.03997			-0.60851			-1.34029		
Usual		0.32906	-0.121		0.26525	-2.294		0.40878	-3.279
Robust					0.20562	-2.959		0.38732	-3.460
Correct		0.32922	-0.121		0.21042	-2.892		0.44897	-2.985
LM t-Stat			0.488			2.729			2.451
1980	-0.25644			-0.83008			-1.55812		
Usual		0.22366	-1.147		0.23867	-3.478		0.34194	-4.557
Robust					0.20367	-4.076		0.37494	-4.156
Correct		0.22514	-1.139		0.22061	-3.763		0.47715	-3.265
LM t-Stat			2.021			3.428			3.028

	BC			NLS			WNLS		
		Std.Err.	t-stat		Std.Err.	t-stat		Std.Err.	t-stat
1981	-0.26231			-0.83348			-1.55279		
Usual		0.23949	-1.095		0.23998	-3.473		0.34355	-4.520
Robust					0.20180	-4.130		0.37413	-4.150
Correct		0.24145	-1.086		0.21887	-3.808		0.47445	-3.273
LM t-Stat			2.425			3.641			3.337
1982	-0.23864			-0.86281			-1.62734		
Usual		0.41105	-0.581		0.26492	-3.257		0.36260	-4.488
Robust					0.21198	-4.070		0.38794	-4.195
Correct		0.41200	-0.579		0.22901	-3.768		0.49312	-3.300
LM t-Stat			2.261			2.967			2.707
1983	-0.52874			-1.03206			-1.64489		
Usual		0.28956	-1.862		0.25884	-3.987		0.34384	-4.784
Robust					0.20612	-5.007		0.36911	-4.456
Correct		0.29058	-1.820		0.22967	-4.494		0.48598	-3.385
LM t-Stat			3.669			1.806			1.375
1984	-0.62799			-1.07003			-1.64515		
Usual		0.29118	-2.157		0.24737	-4.326		0.34112	-4.823
Robust					0.21541	-4.967		0.36877	-4.461
Correct		0.29362	-2.139		0.23952	-4.467		0.48568	-3.387
LM t-Stat			4.522			2.051			1.484

--- Table 4 ---
Alternative Estimates of Standard Errors and t-Statistics
Chow Data

	BC			NLS			WNLS		
		Std.Err.	t-stat		Std.Err.	t-stat		Std.Err.	t-stat
Lambda	0.12913			-2.02749			-3.05296		
Usual		0.14056	0.919		0.43165	-4.697		0.23361	-13.069
Robust					0.40761	-4.974		0.24741	-12.340
Correct		0.14167	0.911		0.40761	-4.974		0.24741	-12.340
LM t-Stat			1.479			1.438			3.418
Constant	0.04120			0.68441			0.36960		
Usual		0.24733	0.167		0.19839	3.450		0.02142	17.253
Robust					0.17167	3.987		0.02788	13.256
Correct		0.30742	0.134		0.16964	4.034		0.03127	11.820
LM t-Stat			0.128			2.012			1.245
MULT	-0.00001			-0.00094			-0.00125		
Usual		9.82E-06	-0.947		0.00052	-1.804		0.00044	-2.813
Robust					0.00043	-2.171		0.00048	-2.604
Correct		1.00E-05	-0.870		0.00045	-2.084		0.00049	-2.535
LM t-Stat			1.524			0.076			0.112
MEM	0.00042			1.46E-05			9.52E-06		
Usual		7.10E-05	5.894		6.95E-06	2.106		3.57E-06	2.668
Robust					5.98E-06	2.446		3.10E-06	3.071
Correct		9.00E-05	4.824		5.69E-06	2.570		3.47E-06	2.744
LM t-Stat			5.531			2.894			3.968
ACCESS	-0.00008			-0.04812			-0.00247		
Usual		4.74E-05	-1.694		0.02594	-1.855		0.00154	-1.606
Robust					0.02155	-2.233		0.00168	-1.472
Correct		5.00E-05	-1.679		0.02150	-2.238		0.00209	-1.181
LM t-Stat			2.040			2.406			2.767
1961	-0.15710			-0.29943			-0.16786		
Usual		0.38903	-0.404		0.14187	-2.111		0.05856	-2.867
Robust					0.12095	-2.476		0.05497	-3.054
Correct		0.38590	-0.407		0.11645	-2.571		0.06123	-2.742
LM t-Stat			0.491			0.304			0.199
1962	0.09273			-0.15963			-0.04633		
Usual		0.29238	0.317		0.08247	-1.936		0.01608	-2.881
Robust					0.06872	-2.323		0.01452	-3.190
Correct		0.30276	0.306		0.06821	-2.340		0.01703	-2.720
LM t-Stat			0.218			1.162			[illegible]
1963	-0.25717			-0.21562			-0.09324		
Usual		0.30548	-0.842		0.10775	-2.001		0.03105	-3.003
Robust					0.09052	-2.382		0.02990	-3.118
Correct		0.30768	-0.836		0.08889	-2.426		0.03403	-2.740
LM t-Stat			0.704			1.859			0.777
1964	-0.14770			-0.22731			-0.08507		
Usual		0.35770	-0.413		0.11520	-1.973		0.02779	-3.061
Robust					0.09655	-2.354		0.02647	-3.214
Correct		0.35788	-0.413		0.09526	-2.386		0.03098	-2.746
LM t-Stat			0.480			2.038			0.868
1965	-0.73291			-0.23831			-0.08294		
Usual		0.31279	-2.343		0.12272	-1.942		0.02742	-3.025
Robust					0.10301	-2.314		0.02653	-3.126
Correct		0.31540	-2.324		0.10206	-2.335		0.03021	-2.745
LM t-Stat			2.089			2.048			2.890

towards the end of the sample in 1982-84, LM t-statistics on the WNLS time dummies become smaller (in absolute value) than those based on the other procedures.[9]

Finally, another interesting finding in Table 3 concerns alternative estimates of the standard errors. For the BC procedure, traditional and "correct" standard error estimates are quite similar, with no discernible inequality relationship between them occurring. For the NLS and WNLS procedures, we see that correcting for the sample geometric mean has a somewhat larger effect than in the BC case, but most of this is due to the heteroskedasticity adjustment implicit in the "correct" standard errors.

The final data set we use in comparing the BC, NLS and WNLS procedures is that underlying the classic study by Gregory Chow on estimating the prices and price elasticity of demand for computers. Our results from the CHOW data set are presented in Table 4. A number of results are worth noting.

First, estimates of λ vary dramatically across estimation procedures. While the 0.129 BC estimate of λ is positive, small and statistically insignificant, the NLS (-2.03) and especially the WNLS (-3.05) estimates are negative, large and statistically significant. A negative estimate of λ in the computer market is not entirely unexpected (see Jack E. Triplett (1989) for a conjecture that this might occur), and demonstrates the importance of allowing λ to vary outside the (0,1) interval when implementing such models empirically. For the estimated slope coefficients on the MULT, MEM and ACC variables, no sign differences occur among the three estimation

[9]Implications of these results for price index computation are discussed in detail by Berndt, Showalter and Wooldridge (1990).

procedures, but one sign variation is found among the annual time dummy variables--that for 1962.

In terms of standard error estimation procedures, for BC the traditional and "correct" method yield roughly similar results, with the "correct" estimates being slightly larger than the traditional in all cases but one (the 1961 time dummy). Interestingly, for NLS the corrections (both heteroskedasticity and sample geometric mean) actually lower the standard errors for all the parameters. For WNLS we have mixed results. Generally the robust standard errors are smaller than the traditional ones while the "correct" standard errors are, with one exception, the highest of the three estimates (the exception being the MEM coefficient where the "correct" standard error is higher than the traditional but lower than the robust).

On the basis of parameter estimates and inference, therefore, we conclude that substantial differences are found among the BC, NLS and WNLS estimates, particularly with the COLE and CHOW data sets. We see no systematic effect of adjusting for the sample geometric mean, although there is some evidence that using White's robust standard errors might be an adequate approximation for the "correct" standard errors.

We now move on to a comparison of estimation methods using goodness-of-fit criteria. Recall from our earlier discussion that if one defines the residual as $\hat{e}_t \equiv y_t - f(x_t, \hat{\beta}, \hat{\lambda})$, then by construction the NLS estimator will always produce a lower sum of squared residuals than the WNLS method. However, we cannot say that NLS will result in a lower sum of squared residuals than BC when the residuals are calculated using equations (2.7)-(2.9) (although we suspect that this will generally be the case). As a result, it is important to use criteria other than sums of squared residuals when comparing the BC, NLS and WNLS procedures.

Table 5 - CPS 1978 Data

Summary Statistics

	Y	IBC	BC	NLS	WNLS
Min	0.116	0.477	0.516	0.532	0.542
Max	5.314	2.129	2.268	2.546	2.566
25%	0.698	0.816	0.878	0.859	0.857
75%	1.396	1.239	1.327	1.335	1.330
Std Dev	0.607	0.314	0.334	0.362	0.361
Mean	1.129	1.050	1.126	1.129	1.129

Sum of Squared Residuals

IBC	-- 135.64
BC	-- 131.49
NLS	-- 130.35
WNLS	-- 130.40

Correlation Matrices

--Centered--

Y	1.000				
IBC	0.592	1.000			
BC	0.592	1.000	1.000		
NLS	0.596	0.995	0.995	1.000	
WNLS	0.595	0.995	0.995	1.000	1.000

--Uncentered--

Y	1.000				
IBC	0.924	1.000			
BC	0.924	1.000	1.000		
NLS	0.925	0.999	0.999	1.000	
WNLS	0.925	0.999	0.999	1.000	1.000

--Residuals--

IBC	1.000			
BC	0.999	1.000		
NLS	0.993	0.996	1.000	
WNLS	0.993	0.996	1.000	1.000

Table 6 - Cole Data

Summary Statistics

	Y	IBC	BC	NLS	WNLS
Min	0.294	0.573	0.583	0.296	0.312
Max	2.766	2.755	2.757	2.439	2.128
25%	0.742	0.761	0.767	0.772	0.817
75%	1.291	1.178	1.182	1.285	1.377
Std Dev	0.565	0.501	0.500	0.510	0.465
Mean	1.132	1.126	1.131	1.132	1.113

Sum of Squared Residuals

IBC	-- 6.70
BC	-- 6.71
NLS	-- 5.24
WNLS	-- 6.02

Correlation Matrices

--Centered--

Y	1.000				
IBC	0.876	1.000			
BC	0.875	1.000	1.000		
NLS	0.904	0.968	0.968	1.000	
WNLS	0.892	0.894	0.893	0.969	1.000

--Uncentered--

1.000				
0.977	1.000			
0.977	1.000	1.000		
0.982	0.995	0.995	1.000	
0.979	0.983	0.983	0.995	1.000

--Residuals--

IBC	1.000			
BC	1.000	1.000		
NLS	0.882	0.881	1.000	
WNLS	0.641	0.641	0.868	1.000

Table 7 - Chow Data

Summary Statistics

	Y	IBC	BC	NLS	WNLS
Min	0.090	0.055	0.091	0.021	0.086
Max	21.218	74.892	81.636	21.218	21.221
25%	0.398	0.810	1.057	0.610	0.853
75%	2.254	1.299	1.646	2.019	1.889
Std Dev	3.003	8.215	8.944	2.844	2.780
Mean	2.034	2.170	2.574	1.999	2.029

Sum of Squared Residuals

IBC	--	3089.09
BC	--	3829.93
NLS	--	87.08
WNLS	--	91.47

Correlation Matrices

--Centered--

Y	1.000				
IBC	0.778	1.000			
BC	0.782	1.000	1.000		
NLS	0.939	0.813	0.816	1.000	
WNLS	0.935	0.827	0.831	0.997	1.000

--Uncentered--

Y	1.000				
IBC	0.766	1.000			
BC	0.778	1.000	1.000		
NLS	0.958	0.790	0.801	1.000	
WNLS	0.956	0.797	0.808	0.998	1.000

--Residuals--

IBC	1.000			
BC	0.999	1.000		
NLS	0.125	0.107	1.000	
WNLS	0.139	0.117	0.977	1.000

In Table 5 we present summary statistics (min, max, 25%, 75%, std. dev. and mean) for the observed (scaled by the sample geometric mean) dependent variable, y, and fitted values, where the latter are computed in four ways, as discussed in Section 2: IBC (Incorrect Box-Cox, Box-Cox using equation (2.5)), BC (Box-Cox using equation (2.7)-(2.9)), NLS (nonlinear least squares fitted value) and WNLS (weighted nonlinear least squares fitted values), all for the CPS78 data. Corresponding summary statistics for the COLE and CHOW data sets are given in Tables 6 and 7.

As seen in Table 5, for the CPS78 data the distribution of the fitted values is roughly similar for all four procedures, although the mean of the fitted value for IBC is about 10% less than the sample mean of y, implying that for the IBC, the mean "residual" is non-zero. For the COLE data (see Table 6), while mean fitted values are all approximately equal and close to the sample observed mean, the IBC and BC minimum fitted values (0.573 and 0.583) are substantially larger than those for NLS (0.296) and WNLS (0.312), and for the sample observed min (0.294). However, maximum fitted values of the IBC (2.755) and BC (2.757) are close to the observed sample max (2.766), but these maximum values are larger than the maximum fitted values based on the NLS (2.439) and WNLS (2.128) procedures.

Finally, for the CHOW data (see Table 7), greater diversity appears. The mean fitted values for IBC (2.170), NLS (1.999) and WNLS (2.029) are quite close to the sample mean of the observed y (2.034), but the mean fitted value from BC (2.574) is about 20% larger. Although the min (0.086) and max (21.221) fitted values from WNLS are virtually identical to those observed (0.090 and 21.218), the min fitted values for NLS (0.021) and IBC (0.055) are smaller, and the max fitted values for IBC (74.892) and BC (81.636) are much larger than for observed y (21.218). Since the max values for IBC and BC are so much larger than for the observed y, several IBC and BC residuals will be correspondingly large, one might expect that the sum of squared residuals will be correspondingly large, and therefore that the sum of squared residuals based on the BC and IBC methods and the CHOW data will be much larger than for the NLS and WNLS methods. This is in fact what occurs; as seen in Table 7, with the CHOW data the sums of squared residuals for IBC (3089.09) and BC (3829.93) are much larger than for WNLS (91.47) or NLS (87.08). For the CPS78 data (Table 5), differences in the sums of squared residuals are very small, while for the COLE data (Table 6), the differences are only slightly larger.

In the middle panels of Table 5,6 and 7, we present simple correlations between the observed and four sets of fitted values (IBC, BC, NLS and WNLS), both centered about their sample means and uncentered. For the CPS78 data (Table 5), the uncentered correlations are all very large (above 0.9), and for the centered correlations with observed y, the fitted value correlations are all very similar (about 0.59). For the COLE data (Table 6), simple correlations display a bit more diversity, but differences are not dramatic. With the CHOW data, however (Table 7), two groups of correlations differ. While the IBC and BC fitted values are very highly correlated (the centered and uncentered correlation are each 1.000) with each other, and while the NLS and WNLS reveal similarly high correlations (0.997 centered, 0.998 uncentered), simple correlations between the IBC, BC and NLS-WNLS fitted values are lower, around 0.8 for both the centered and uncentered data.

These correlations among fitted values and between fitted and actual values of y imply correlation structures among residuals. Simple (uncentered) residual correlations for the CPS78, COLE and CHOW data sets are presented in the bottom panel of Tables 5, 6 and 7, respectively. As seen in Table 5, the inter-correlations among the IBC, BC, NLS and WNLS residual for the CPS78 data set are all very high--greater than 0.99. For the COLE data set (Table 6), we find that the WNLS residuals have a relatively low correlation with the BC-IBC residuals (0.641), with the remaining correlations 0.86 or above.set (Table 7), however, three clusters of correlations become evident. While correlations between IBC and BC residuals (0.999) and between NLS and WNLS residuals (0.977) remain very high, simple correlations between one of IBC-BC and one of NLS-WNLS are very low--between about 0.11 and 0.13. With the CHOW data, therefore, two very distinct groups of residuals emerge--one set based on Box-Cox variants, and the other on nonlinear least squares variants. For this data set in particular, the transformation and nonlinear least squares methods yield very different results.

5. Concluding Remarks

Our purpose in this paper has been to compare empirically two distinct approaches to choosing a functional form--the Box-Cox and nonlinear least squares procedures--based on three publicly available data sets.

We can summarize our findings as follows. First, we provided a rather persuasive empirical example demonstrating that with both the Box-Cox (BC) and nonlinear least squares (NLS) procedures, while t-statistics on the transformation parameter λ are invariant to arbitrary scaling of the dependent variable, the t-statistics on slope coefficients, intercepts and dummy variables can be changed dramatically simply by arbitrarily re-scaling the data. We also noted that since the t-statistic is a Wald test statistic, this lack of invariance is not surprising, and we eliminated it by employing a computationally more cumbersome Lagrange multiplier test statistic, systematically excluding one variable at a time and re-estimating. We conclude, therefore, that while in practice in these nonlinear models scaling issues are very important, they can be resolved though use of the LM test statistic procedure. Future research that focuses on necessary and sufficient conditions for such lack of scaling invariance, as well as on more computationally efficient ways of doing scale-invariant testing of exclusionary restrictions, would appear to be most useful.

Second, we have found that differences among the BC, NLS and weighted nonlinear least squares (WNLS) parameter estimates vary by data set, and that little in general can be stated concerning what a researcher should expect with a particular data set. Specifically, in one data set (CPS78) parameter estimates differed very little among alternative estimators, in a second data set (COLE) the differences were substantially larger--sometimes even resulting in different signs for estimated coefficients, and in our third data set (CHOW) the differences were very large, with the estimated transformation parameter λ having a different sign depending on the estimation procedure employed. Since in some cases we find substantial differences among estimators, we are now faced with issues assessing which estimator is "best" in terms of yielding estimates closest to the "true" parameters. Our results therefore imply that empirical assessment of these alternative estimators based on a well-designed Monte-Carlo approach is warranted.[10]

[10]Research on this topic is currently underway. See Showalter (1990).

Third, we have computed standard error estimates using traditional, heteroskedasticity-robust and simultaneous heteroskedasticity-robust and sample geometric mean-adjusted computational procedures. Our results suggest that when there are differences among these alternative standard error estimates, most of the difference can be attributed to adjusting for heteroskedasticity; the marginal change induced by adjusting for the random sample geometric mean of the dependent variable is relatively minor.

Fourth, in terms of fitted values and residuals, we have found that in some cases the common but incorrect Box-Cox (IBC) and correct Box-Cox(BC) procedures yield fitted values much greater than (less than) the sample maximum (minimum) values of the observed y, and that in such cases the resulting extremely large residuals for IBC and BC yield very large sums of squared residuals, much larger than that for NLS and WNLS. In these cases, while the correlations between IBC and BC residuals, and between NLS and WNLS residuals, are very high, the IBC-BC and NLS-WNLS residuals tend to cluster in two distinct groups, with simple correlations between any one of IBC-BC and one of NLS-WNLS being very small (less than 0.15). Which of these residuals are more "correct" depends of course on the true parameters and model, and in this study those are still unknown. Further research on this topic using Monte Carlo approaches would be useful

Finally, in this paper we have reported results using only the Box-Cox transformation on the dependent variable, and have employed "natural" (i.e., untransformed) values for the explanatory variables. In particular, we have not reported results when some explanatory variables are transformed into logarithms (as was done in the original COLE and CHOW studies) or are transformed using the Box-Tidwell procedures. We have done some research on these issues, however, and can briefly report that when one employs logarithmic transformations of explanatory variables as was done in the original studies by Cole et al. and Chow, differences among the BC, NLS and WNLS estimates of λ become rather small, and typically our λ estimates were insignificantly different from zero, thereby lending support to the log-log functional form specification used by Cole et al. and Chow.[11] However, when Box-Tidwell-type procedures are employed, differences among the various estimation procedures re-emerge.

[11]For further discussion, see Berndt, Showalter and Wooldridge (1990).

References

Amemiya, T. and J.L. Powell (1981), "A Comparison of the Box-Cox Maximum Likelihood Estimator and the Non-Linear Two Stage Least Squares Estimator," *Journal of Econometrics*, 17, 351-381.

Berndt, E.R. (1990), *The Practice of Econometrics: Classic and Contemporary*, Reading, MA: Addison-Wesley Publishers, Forthcoming.

Berndt, E.R., B.H. Hall, R.E. Hall, and J.A. Hausman (1974), "Estimation and inference in nonlinear structural models," *Annals of Economic and Social Measurement*, 3, 653-665.

Berndt, E.R., M.H. Showalter and J.M. Wooldridge (1990), "Hedonic Price Indexes for Computers and Peripherals: Box-Cox Formulations and a Nonlinear Least Squares Alternative," Massachusetts Institute of Technology, mimeo, July.

Bollershev, T. and J.M. Wooldridge (1988), "Quasi-maximum Likelihood Estimation of Dynamic Models with Time Varying Covariances," Working Paper 505, Department of Economics, Massachusetts Institute of Technology.

Box, G.E.P. and D.R. Cox (1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society*, Series B, 26, 211-252.

Chow, G.C. (1967), "Technological Change and the Demand for Computers," *American Economic Review*, 57, 1117-1130.

Cole, R., Y.C. Chen, J.A. Barquin-Stolleman, E. Dulberger, N. Helvacian and J.H. Hodge (1986), "Quality-Adjusted Price Indexes form Computer Processors and Selected Peripheral Equipment," *Survey of Current Business*, 66, 41-50.

Dagenais, M.G. and J.-M. Dufour (1986), "Invariance, Nonlinear Models and Asymptotic Tests," Working Paper 3287, Department of Economics, University of Montreal, December.

Draper, N.R. and D.R. Cox (1969), "On Distributions and Their Transformation to Normality," *Journal of the Royal Statistical Society*, Series B, 31, 472-476.

Hinkley D.V. and G. Runger (1984), "The Analysis of Transformed Data," *Journal of the American Statistical Association*, 79, 302-309.

Khazzoom, J.D. (1989), "A Note on the Application of the Nonlinear Two-Stage Least-Squares Estimator to a Box-Cox-Transformed Model," *Journal of Econometrics*, 42, 377-379.

Neyman, J. (1959), "Optimal Asymptotic Tests of Composite Statistical Hypotheses," in U. Grenander, ed., *Probability and Statistics*, Almqvist and Wiksell: Stockholm, 213-234.

Poirier, D.J. and A. Melino (1978), "A Note on the Interpretation of Regression Coefficients Within a Class of Truncated Distributions," *Econometrica*, 46, 1207-1209.

Seakes, T.G. and S.K. Layson (1983), "Box-Cox Estimation with Standard Econometric Problems," *Review of Economics and Statistics*, 65, 160-164.

Showalter, M.H. (1990), "A Monte Carlo Investigation of the Box-Cox Model and a Nonlinear Least Squares Alternative", MIT mimeo.

Spitzer, J.J. (1982), "A Primer on Box-Cox Estimation," *Review of Economics and Statistics*, 64, 307-313.

Spitzer, J.J. (1984), "Variance Estimates in Models with Box-Cox Transformations: Implications for Estimation and Hypothesis Testing," *Review of Economics and Statistics*, 66, 645-652.

Triplett, J.E. (1989), "Price and Technological Change in a Capital Good: A Survey of Research on Computers," Chapter 4 in Dale W. Jorgenson and Ralph Landau, eds., *Technology and Capital Formation*, Cambridge, MA: MIT Press, 127-213.

White, H. (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817-838.

Wooldridge, J.M. (1990), "Some Alternatives to the Box-Cox Regression Model," mimeo, MIT Department of Economics, January.

Zarembka, P. (1968), "Functional Form in the Demand for Money," *Journal of the American Statistical Association*, 63, 502-511.